Outcomes Monitoring and the Testing of New Psychiatric Treatments: Work Therapy in the Treatment of Chronic Post-Traumatic Stress Disorder

Robert Rosenheck, Marilyn Stolar, and Alan Fontana

Objective. To evaluate the effectiveness of a work therapy intervention, the Department of Veterans Affairs (VA) Compensated Work Therapy program (CWT), in the treatment of patients suffering from chronic war-related post-traumatic stress disorder (PTSD); and to demonstrate methods for using outcomes monitoring data to screen previously untested treatments.

Data Sources/Study Setting. Baseline and four-month follow-up questionnaires administered to 3,076 veterans treated in 52 specialized VA inpatient programs for treatment of PTSD at facilities that also had CWT programs. Altogether 78 (2.5 percent) of these patients participated in CWT during the four months after discharge.

Study Design. The study used a pre-post nonequivalent control group design.

Data Collection/Extraction Methods. Questionnaires documented PTSD symptoms, violent behavior, alcohol and drug use, employment status, and medical status at the time of program entry and four-months after discharge from the hospital to the community. Administrative databases were used to identify participants in the CWT program. Propensity scores were used to match CWT participants and other patients, and hierarchical linear modeling was used to evaluate differences in outcomes between treatment groups on seven outcomes.

Principal Findings. The propensity scaling method created groups that were not significantly different on any measure. No greater improvement was observed among CWT participants than among other patients on any of seven outcome measures.

Conclusions. Substantively this study suggests that work therapy, as currently practiced in VA, is not an effective intervention, at least in the short term, for chronic, war-related PTSD. Methodologically it illustrates the use of outcomes monitoring data to screen previously untested treatments and the use of propensity scoring and hierarchical linear modeling to adjust for selection biases in observational studies.

Key Words. Outcomes, quality, post-traumatic stress disorder, propensity scores

OUTCOMES MANAGEMENT AND TECHNOLOGY ASSESSMENT

Over a decade ago, outcomes management in healthcare was described (Ellwood 1988) and heralded as the "third revolution" in twentieth-century medical care (Relman 1988). Since that time, it has been widely accepted that quality in healthcare cannot be adequately assessed by traditional indicators such as professional licensure, continuing medical education, focused chart reviews, or even adherence to recommended clinical pathways (Brennan and Berwick 1996; Epstein 1990). Meaningful assessment of the value of healthcare services requires, in addition, the evaluation of clinical outcomes assessed by direct, systematic measurement of the results of treatment (Blumenthal 1996; Millenson 1997; Sederer, Dickey, and Hermann 1996).

In one of the earliest explications of the outcomes management imperative, Ellwood (1988:1552) stated that "outcomes management's closest relative is the clinical trial." He suggested that outcomes monitoring data would eventually be used to compare existing treatments and to evaluate new technologies, thereby avoiding both the expense of costly clinical trials and the loss of generalizability that was often entailed in selective recruitment for such trials. Ellwood envisioned the construction of comprehensive national outcome databases for specific diseases and the use of those databases to test promising treatments. Such treatments could be tried initially on a small number of patients. These patients could be matched with similar patients in a large database to allow assessment of the relative effectiveness of the new treatment.

Although outcomes monitoring has been embraced as the desired standard for quality assessment, the practical difficulties and costs of implementing such extensive data collection have severely limited its deployment in actual practice (Steinwachs, Wu, and Skinner 1994; Clardy, Booth,

Address correspondence to Robert Rosenheck, M.D./182, Director, Northeast Program Evaluation Center, VA Connecticut Healthcare System, 950 Campbell Avenue, West Haven CT 06516. Dr. Rosenheck is also Professor of Psychiatry and Public Health, Yale University Department of Psychiatry, School of Epidemiology and Public Health; and Director, Evaluation Division, National Center for PTSD, Yale University School of Medicine. Marilyn Stolar, M.A. is a Biostatistician, NEPEC, VAMC West Haven, CT and Yale University School of Epidemiology and Public Health; and Alan Fontana, Ph.D. is Associate Director, NEPEC, VAMC West Haven, CT, Yale University Department of Psychiatry, and Evaluation Division, National Center for PTSD. This article, submitted to *Health Services Research* on August 25, 1998, was revised and accepted for publication on January 15, 1999.

Smith, et al. 1998). Furthermore, Ellwood and others did not address the problem of potentially confounding differences in the types of patients who receive different treatments in general practice—the very reason why the randomized clinical trial remains the gold standard for evaluating clinical efficacy. Thus, even though the use of outcomes monitoring data to test new treatments is promising in theory, we are aware of no previous studies in which this approach has been applied in the field of behavioral healthcare.

WAR-RELATED POST-TRAUMATIC STRESS DISORDER (PTSD)

The treatment of war-related post-traumatic stress disorder (PTSD) is a major priority for the Department of Veterans Affairs (VA) healthcare system. National survey data indicate that almost 500,000 veterans of the Vietnam era meet minimal diagnostic criteria for PTSD (Kulka, Schlenger, Fairbank, et al. 1990), and 80,000 veterans suffering from debilitating problems such as nightmares, flashbacks, and profound social withdrawal seek help for war-related PTSD from VA each year (Rosenheck et al. 1997). Outcome studies suggest that conventional psychosocial and pharmacotherapeutic treatments have limited efficacy, especially in severe and persistent cases of PTSD (Fontana and Rosenheck 1997a; Rosenheck and Fontana 1996; Solomon, Gerrity, and Muff 1992; Davidson 1997). It is now over two decades since the last U.S. soldier left Vietnam and the psychiatric sequelae of the Vietnam war are, by definition, chronic and, in most cases that present to the VA, seriously disabling (Rosenheck and Fontana 1996).

It has been suggested that rehabilitative treatments of the type developed specifically for patients with severe and persistent mental illnesses (Mueser, Drake, and Bond 1997; Lehman 1995) may be effective for veterans with chronic PTSD (Rosenheck and Fontana 1996; Friedman and Rosenheck 1996). Conventional psychotherapies for PTSD focus on cathartic recollection, progressive desensitization, deconditioning, or cognitive reframing of traumatic memories. However, these approaches seem to have limited effectiveness in chronic war-related PTSD of the type currently experienced by Vietnam combat veterans (Solomon, Gerrity, and Muff 1992). The problems of such veterans are quite different from those of people with recent traumatic exposure because, in most cases, their difficulties have persisted for over two decades and many have already had extensive psychotherapeutic treatment. Recent outcome studies suggest that intensive, prolonged exploratory treatment may even result in worsening of symptoms in severe and persistent

PTSD (Johnson, Rosenheck, Fontana, et al. 1996; Johnson et al. 1997). In such cases, once the exploration of traumatic memories has been prolonged without relief, continued therapeutic exposure may reinforce rather than extinguish painful memories, since such exposure may become associated with the supportive environment of treatment or the camaraderie of group therapy. Work therapy may be more effective because it rewards behavior that is unconnected to war memories, provides daily structure, develops functional skills, and offers a source of pride and purpose, all of which may help refocus the patient's attention on positive aspects of the present rather than on darker aspects of the past. No studies, however, have yet been conducted to evaluate the effectiveness of rehabilitation-oriented treatment of chronic war-related PTSD.

OUTCOMES MONITORING OF PTSD TREATMENT

In 1993, VA initiated development of a multi-component National Mental Health Program Performance Monitoring system to evaluate the behavioral healthcare provided to over 600,000 patients per year at over 150 medical centers across the country (Rosenheck 1996). This performance monitoring system is based primarily on administrative data of the HEDIS type (National Committee on Quality Assurance [NCQA] 1995) and an annual systemwide patient satisfaction survey (Rosenheck, Wilson, and Meterko 1997). A special subcomponent of the system, however, collects outcome interview data, at the time of program entry and four months after discharge, among patients admitted to specialized intensive programs that treat war-related PTSD (Fontana and Rosenheck 1997b).

In this study we use data from over 6,000 participants in the PTSD outcomes monitoring effort to evaluate a rehabilitation-oriented work therapy in the treatment of war-related PTSD. VA's Compensated Work Therapy (CWT) program is a therapeutic work-for-pay program, operating at almost 100 VA medical centers across the country (Seibyl et al. 1997). Through this program, private businesses or federal agencies enter a contract with VA for work that is performed by patients in supervised rehabilitation programs.

Our objectives in this study are both methodological and substantive. First, we seek to evaluate the potential of a rarely used treatment for a severe condition, difficult to treat, in which an argument can be made on conceptual grounds for the treatment's effectiveness. At the same time, we illustrate the application of analytic methods that have specific relevance for using outcomes monitoring data to evaluate promising psychiatric treatments for which for further development and further evaluation may be appropriate.

METHODS

A special subcomponent of the National Mental Health Program Performance Monitoring system was designed to monitor clinical outcomes from inpatient and residential programs that provide specialized treatment for veterans with war-related post-traumatic stress disorder (Fontana and Rosenheck 1997b). These programs were designated for ongoing, intensive performance evaluation because of the high priority placed on treatment of war-related PTSD in VA, and because these programs are characterized by high treatment intensity and high cost (Fontana and Rosenheck 1997b). Sixty-two programs participated in this evaluation effort from March 1993 through February 1996, but only 52 were located at medical centers that also provided CWT treatment.

Patients admitted to these programs were assessed with a brief standardized self-report questionnaire at the time of admission and again four months after discharge. These questionnaires were administered either in face-to-face encounters or, when necessary, over the telephone.

Compensated Work Therapy

Outcomes monitoring data were merged with national computerized VA outpatient workload files in which delivery of all VA outpatient services was documented, and the merged data were used to identify veterans who participated in VA's CWT (Seibyl et al. 1997) program during the first four months after discharge from a specialized inpatient PTSD program.

Sample

Between March 1, 1993 and February 29, 1996, 5,065 veterans were enrolled in the monitoring protocol at sites with CWT programs and successfully merged VA workload files, and 3,076 (61 percent) of these veterans were successfully contacted after discharge and completed the follow-up questionnaire. Comparison of baseline characteristics of the veterans who were successfully followed up and those who were not, using logistic regression, showed that veterans who were followed up were older, were more likely to be white, were more likely to be married, had been diagnosed with more severe medical problems at the time of program entry, had better employment histories, and had more years of education. No differences were found in PTSD symptoms or substance abuse.

Of 128 veterans who participated in CWT, 78 were successfully followed up (61 percent). These veterans had an average of 25.8 days of CWT treatment (s.d. = 22.0, median = 21) during the four-month period, for an average of 1.6 days of participation per week. In general, participants in CWT work in either piecework programs in VA workshops or in community-based supported employment placements. Wages in both cases are generated through contracts with private businesses. National monitoring data show that CWT participants at the sites involved in this study worked an average of 25.4 hours per week, earned an average of $4.55 per hour, and participated in the program for an average of 4.6 months (Seibyl et al. 1997).

Measures

Sociodemographic Characteristics. Sociodemographic data obtained at baseline included measures of age, race, marital status, education, history of incarceration, current employment, and receipt of VA compensation for PTSD.

Clinical Outcome Measures. Clinical outcomes that were assessed included (1) PTSD symptoms, (2) substance abuse, (3) violent behavior, (4) employment, and (5) medical status.

Because of their particular significance for specialized PTSD programs, PTSD symptoms were measured in two ways, using (1) the Short Form of the Mississippi Scale for Combat-Related PTSD (range = 0–44), an instrument that has been validated in a large sample of outpatients (Fontana and Rosenheck 1994), and (2) a four-item PTSD Scale (range 0–16) developed at the Northeast Program Evaluation Center (the NEPEC PTSD scale) (Cronbach alpha = 0.67). The NEPEC PTSD Scale correlated 0.61 and 0.74 with the Short Mississippi Scale at admission and at four months follow-up, respectively. Thus, the NEPEC PTSD Scale and the Short Mississippi Scale correlation was sufficiently large to indicate that the two instruments were measuring the same domain, but not so large that either was redundant.

In an intensive outpatient PTSD study, the NEPEC PTSD Scale and the Short Mississippi Scale correlated 0.63 and 0.64, respectively, with a continuous PTSD score derived from the SCID PTSD module (Structured Clinical Interview for DSM-III) (Fontana, Rosenheck, and Spencer 1993). Additionally, in an outcome study of intensive inpatient treatment of PTSD (Rosenheck et al. 1997), the NEPEC PTSD Scale and the Short Mississippi

Scale correlated .40 and .39, respectively, with the CAPS (Clinician Administered PTSD Scale), a well-validated observer rating scale (Spitzer and Williams 1985; Blake 1994; Weathers and Litz 1994).

Alcohol abuse and drug abuse were measured using the composite indexes from the Addiction Severity Index (ASI) (range 0–1) (McLellan, Luborsky, Cacciola, et al. 1985), a widely used and well-validated measure of substance abuse outcomes. Violent behavior was measured by four items that were adapted from the National Vietnam Veterans Readjustment Study (range 0–4) (Kulka, Schlenger, Fairbank, et al. 1990): (1) destruction of property, (2) threatening someone with physical violence without a weapon, (3) threatening someone with a weapon, and (4) physically fighting with someone (Cronbach alpha = 0.71). Employment and medical status were measured using the relevant composite indexes from the ASI (range 0–1 with high scores representing better employment and worse medical status) (McLellan, Luborsky, Cacciola, et al. 1985).

Statistical Analysis

The major challenge in using observational data to compare treatment results is that participation in selected treatments such as CWT is determined by natural referral processes rather than by random assignment. As a result, outcome differences between groups may reflect either treatment effects or selection biases in treatment choice.

Randomized controlled experiments allow researchers to capitalize on chance to achieve treatment and control assignments that are approximately balanced with respect to the multivariate distribution of subject characteristics that might influence the outcome of interest. This balance allows for assessment of the effect of treatment on group differences in outcome measures under the probabilistic assumption of "all else being about equal."

In observational studies, however, subjects who receive particular treatments are likely to be systematically chosen for those treatments on the basis of attributes that are associated with treatment type and with outcome measures. For example, in our study, patients who were identified as functionally able to partake in and benefit from a work therapy program were chosen to participate. Thus, when treatment and control subjects were followed over time, we could not be sure if improvement in the outcome measures reflected treatment actually received or the effect of being better candidates for CWT. In an ideal scenario, we would be able to observe and compare the outcomes of one group of potential treatment subjects under two different scenarios: without the treatment program and with the treatment program. This would match out all differences between the treatment and control groups except for the treatment itself.

A. Propensity Score Subclassification. Propensity scoring is a widely accepted statistical technique originally introduced by Rosenbaum and Rubin (1983, 1984, 1985) to replace the collection of potentially confounding covariates in an observational study with a single score measuring the propensity to be similar to those subjects in the treatment group. This score is then used to match treatment and comparison subjects and thereby to minimize the effect of selection bias on the results. The full procedure includes three steps.

Step 1: Computing propensity scores. In this study, logistic regression was performed in which the dependent variable was a dichotomous (0,1) indicator of participation in CWT and the predictor variables were potential confounders of the treatment-outcome association (age, race, marital status, education, current employment, receipt of VA compensation for PTSD, self-identified need for help with employment, history of incarceration, type of PTSD program, and the baseline values of the outcome measures exclusive of the one being addressed in that particular analysis [see further on]). From the fitted model, each subject's log odds (logit) of being assigned to CWT was computed based on the linear combination of his respective set of values for the predictors. The logit was then converted to its corresponding probability and this value was used as the subject's *propensity score,* a measure of multivariate similarity to individuals chosen for CWT. Subjects with propensity scores close to one thus have multivariate profiles that are more like those of participants in CWT while those with scores close to zero are less like CWT participants.

Predictors of participation in CWT. As noted earlier, seven logistic regression models of participation in CWT were used to generate propensity scores with which to construct matched comparison groups appropriate for the evaluation of change on each outcome variable. Results of these seven logistic regression models were quite similar. For example, in the model that excluded the NEPEC PTSD Scale (a measure that was itself not significantly related to participation in CWT), participation in CWT was significantly associated with lower ASI employment problem scores (Wald $\chi^2 = 9.8$, $p = .002$); not being married (Wald $\chi^2 = 11.2$, $p = .0008$); not receiving

VA compensation for PTSD (Wald $\chi^2 = 13.8$, $p < .0002$); wanting help with employment problems at admission (Wald $\chi^2 = 5.6$, $p = .02$).

Step 2: Partitioning into propensity subclasses. The 78 CWT patients were then divided into five equal-size groups on the basis of the quintiles of their propensity scores. The same cutpoints were then used to partition the controls. Every treatment and control subject was now assigned to one of five propensity subclasses. However, because the five groups of controls were of unequal size, one further step in the matching process was necessary.

Step 3: Matching within propensity subclasses. Within the propensity subclasses of the controls, simple random sampling was carried out to achieve a constant M:1 matching of controls to treatment subjects within each propensity group. We used 6:1 matching, since that was the ratio of controls to treatment subjects in the smallest propensity subclass. Thus, 542 subjects (78 treatment, 464 control) were distributed among five equal-size propensity subgroups consisting of 14.3 percent treatment subjects and 85.7 percent control subjects after sampling. This is our analytic sample. Matching within the five propensity subclasses standardizes the propensity distribution of the controls in the analytic sample to that of the CWT group. Calculating the marginal treatment effect is simplified mathematically when there are equal numbers of treatment subjects per subgroup in the analytic sample.

B. Hierarchical Linear Modeling to Adjust for the Effect of Baseline Score on the Magnitude of Treatment Effects. Each of seven outcome variables was measured on two occasions: at baseline and at four months after discharge from the hospital. The effect of CWT was estimated via separate models for each outcome. We describe the analysis of the Short Mississippi Scale as an example. Note that the baseline value of each particular outcome variable was not used as a predictor to produce propensity scores, but is incorporated into the model of change. Thus, for the analysis of the Short Mississippi Scale, the propensity scores—and thus the particular analytic sample—were defined using all of the aforementioned baseline measures except the Short Mississippi Scale itself. Treatment subjects may thus have different baseline scores on average than the controls in the analytic sample, as that has not been "matched out."

To disentangle baseline score from treatment group, a three-level hierarchical linear model (Bryk and Raudenbush 1992) is used to estimate the treatment effect over time. Measurement times are considered nested within subjects, and subjects are clustered within hospitals. Repeated measures of each outcome variable at baseline and at four months are modeled as a linear function of time and treatment, with a compound symmetric error structure (to account for within-subject correlation of scores at different time points) and an unstructured covariance matrix for the random intercept and slope to account for the clustering of patients within hospitals. The intercept in the model corresponds to each subject's baseline value on the particular outcome measure being analyzed and is controlled for the residual effects of the same predictors used to generate the propensity score. The slope in the model corresponds to the change in the measure over time and is controlled for the subject's deviation from the grand mean of baseline scores on that measure. This adjusted slope thus "matches out" the confounded effect of baseline score from the magnitude of the treatment effect over time. Hierarchical modeling was performed using the SAS® MIXED procedure.

We thus evaluate the interaction of participation in CWT and change over time.

RESULTS

Effectiveness of Matching with Propensity Scores

For each of the seven analytic samples (one for each outcome variable), equality of treatment and control group means or percentages was examined via univariate tests for each potential confounder. For continuous covariates, ANOVA F-tests were used to examine significant differences. For categorical covariates, Pearson χ^2 tests were used. None of the tests supported significant differences between treatment and control subjects on the potential confounders in any of the seven analytic samples. The propensity subclass matching scheme fulfilled its intended purpose.

Sample Characteristics and Outcomes

Sociodemographic and clinical characteristics of the analytic (matched) sample of CWT patients and matched controls ($n = 542$) used in the analysis of the Short Mississippi Scale for PTSD are presented in Table 1. Measures were virtually identical for each of the other six analytic samples.

Within this sample significant improvement was observed over time on six of seven outcome measures, including both measures of PTSD, violence,

Table 1: Characteristics of the Analytic Sample (CWT Patients and Matched Controls with Complete Data, $n = 542$)

	Mean	*s.d.*	*N*	%
Age	45.8	4.4		
Race				
White/Other			325	59.9
African American			183	33.8
Hispanic			34	6.3
Married			79	14.6
Education (years)	12.8	1.4		
Dually diagnosed			346	63.8
Medical diagnoses	0.948	1.005		
Compensation for PTSD			101	18.6
Employment				
None			458	84.5
Part-time			33	6.1
Full-time			51	9.4
Incarcerated				61.4
Never			177	32.6
< 2 weeks			123	22.8
≥ 2 weeks			241	44.6
Patient wants help with employment			417	76.9
Baseline measures				
Short Mississippi	40.42	4.94		
PTSD Scale (NEPEC)	17.04	2.22		
Violent behavior	1.70	1.42		
Alcohol problems (ASI)	0.27	0.28		
Drug problems (ASI)	0.12	0.15		
Medical problems (ASI)	0.52	0.35		
Employment (ASI)	0.37	0.26		

alcohol abuse, drug abuse, and employment, and it was highly statistically significant at $p < .0001$) (Table 2). Thus, participation in these specialized PTSD inpatient programs was associated with significant clinical improvement. The only outcome not showing significant improvement was medical status, which one would not expect to be affected by work therapy.

Comparison of Outcomes Across Treatment Conditions

Table 3 shows that no significant differences occurred in the magnitude of improvement between the CWT and comparison groups, from baseline to follow-up, on any of the seven outcome measures.

Table 2: Clinical Improvement (Analytic Sample Only; $n = 542$)

	Baseline	*s.d.*	*4 Months*	*s.d.*	*Diff*	t*	p
PTSD (Short Mississippi)	40.4	5.0	37.9	7.0	−2.51	8.47	< .0001
NEPEC PTSD scale	17.0	2.2	15.8	3.0	−1.20	9.11	< .0001
Violence	1.7	1.4	1.2	1.2	−0.54	7.75	< .0001
Alcohol problem index	0.27	0.28	0.18	0.21	−0.09	6.98	< .0001
Drug problem index	0.12	0.15	0.07	0.10	−0.04	6.89	< .0001
Medical problem index	0.54	0.34	0.52	0.35	−0.03	1.58	= .11
Employment index	0.30	0.26	0.37	0.26	0.07	6.75	< .0001

*Paired *t*-tests.

Table 3: Effect of Compensated Work Therapy on Outcomes in Comparison with Standard Treatment (Matched Analytic Sample; $n = 542$)

	CWT Effect	*Std. Error*	t	*df*	p
PTSD (Short Mississippi)	−0.7	1.2	−0.6	25	.58
NEPEC PTSD scale	−0.7	0.4	−1.6	25	.12
Violence	−0.02	0.20	−0.1	25	.92
Alcohol problem index	−0.01	0.04	−0.3	25	.75
Drug problem index	−0.03	0.02	−1.6	25	.13
Medical problem index	−0.02	0.05	−0.3	25	.74
Employment index	0.04	0.04	1.07	25	.29

DISCUSSION

Work Therapy in the Treatment of Chronic PTSD

In its use of a large outcomes monitoring database this study did not find evidence that work therapy as currently provided in VA is effective as an addition to standard psychiatric treatment for chronic war-related PTSD, at least during the first four months after discharge from an episode of hospitalization. Despite a cogent rationale for the potential effectiveness of CWT, patients who participated in the program showed no greater improvement in PTSD symptoms, alcohol and drug use, violent behavior, employment activity, or medical status than did a matched sample of other veterans.

Methodological Limitations

However, several possible limitations of these data must be addressed. First, the four-month follow-up time period on which we have reported here is relatively brief, and it is possible that additional benefits may be observed with more extensive involvement in CWT. However, it is likely that many CWT participants were still involved in the work therapy program at the time of the follow-up interview, which should have given them an advantage over patients who received unstructured outpatient treatment or no treatment at all.

Second, although the propensity scaling method used here is the most rigorous approach available for minimizing selection biases in observational studies, we cannot be sure that all relevant variables were included in our calculation of propensity scores (McLellan, Luborsky, Cacciola, et al. 1985; Kulka, Schlenger, Fairbank, et al. 1990; Rosenbaum and Rubin 1983). It is possible that veterans who participated in CWT had a poorer prognosis on some unmeasured characteristic(s) than other veterans. A definitive evaluation of CWT would require a randomized clinical trial, but the negative findings here do suggest that the CWT model should be modified for use with PTSD patients before such a trial is undertaken.

Third, we have only limited information on the nature of these veterans' participation in the CWT program. Through our use of national performance data on all of VA's CWT programs (Rosenheck, Wilson, and Meterko 1997), we validated the fact that each of the hospitals in which CWT participation was recorded has a well-functioning CWT program, but we lack clinical process data on the participation of these veterans in CWT during the period in question or on the specifics of each program. Such data would clarify the nature and quality of the work therapy delivered and could be of value in determining if specific kinds of supported vocational activity are more effective than others.

Several limitations in data quality should also be noted. As in other outcome monitoring efforts (Steinwachs, Wu, and Skinner 1994), the follow-up rate here was relatively low (61 percent), and since patients who were reinterviewed had less severe problems in several areas, the generalizability of our findings is somewhat uncertain. Furthermore, because of the large-scale nature of this project, we relied exclusively on self-report data. Although it is reassuring that, as described in the methods, our PTSD measures were well correlated with a rater-administered instrument, responses in sensitive areas such as violence or substance abuse may have been biased. Finally, although concern has often been expressed that veterans' worries about losing their disability entitlements inhibits candid reporting of their health status, a recent comparison of PTSD outcomes among veterans who were seeking or receiving compensation and veterans who were not showed that this problem is likely to be minimal (Fontana and Rosenheck 1998).

In spite of these limitations, the data presented here suggest that work therapy as currently provided in VA does not offer a short-term therapeutic advantage in the treatment of severe and persistent PTSD, and that this intervention requires further development and refinement if it is to be of value to PTSD patients.

Using Outcomes Monitoring Data to Evaluate Treatment Effectiveness

Support for Clinical Innovation. Perhaps more important than these somewhat disappointing substantive findings is our illustration of the use of a large data set, originally constructed for purposes of performance evaluation, to test the effectiveness of novel or rarely used treatments. Quality of healthcare has been defined by the Institute of Medicine as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes" (Lohr 1990), and healthcare systems are under increasing pressure to generate hard evidence that the services they provide actually improve outcomes. The fact that a treatment such as work therapy "makes sense" in theory, is no longer sufficient in itself to justify its use. However, untested approaches to service delivery, such as work therapy, are also unlikely to be the subject of large-scale clinical trials. There is thus considerable need for observational outcome studies because, to the extent that suitable databases exist, such studies represent an efficient and timely method for testing innovative treatments. There is considerable danger that shrinking budgets, close scrutiny of healthcare expenditures, and standardized disease management algorithms will stifle clinical innovation. Methods for conducting observational outcome studies such as this one are thus urgently needed to support the screening of promising approaches to treatment and service delivery and thereby to encourage creative innovation.

Constructing Appropriate Databases. Two ingredients are crucial to the success of this type of evaluation: the availability of large-scale outcomes databases that allow careful case matching, and the use of appropriate analytic approaches to minimize potentially confounding selection biases. Although VA has long been identified as the largest integrated healthcare system in the United States, the rapid expansion of managed care during the early 1990s has created numerous agencies that have the potential to monitor outcomes on a large scale. To accomplish this objective, as demonstrated in this article, outcomes databases must (1) include large numbers of subjects; (2) use standardized instruments that are appropriate for the clinical condition being treated; (3) measure outcomes in multiple relevant domains; (4) include extensive baseline data in addition to the outcome measures to support matching; (5) collect data at standardized intervals after a sentinel event such as at entry into treatment, admission to the hospital, or discharge from the hospital; and (6) take aggressive steps to achieve the highest possible follow-up rates. Each of these conditions must be met if the data are to allow credible matching and comparison of treatments.

Propensity Scaling. Without appropriate analysis even the best data are of little value. Creating subclasses based on the quintiles of the treatment subjects' propensity scores allows comparison of clinical improvement in a population of treated subjects and in a systematically selected population of control subjects with approximately the same multivariate distribution of sociodemographic, clinical, and service utilization characteristics. In this study we used this approach to create a large, well-matched comparison group for the 78 CWT patients with follow-up data that was not significantly different on any of the available baseline measures.

Propensity scaling has substantial advantages over standard linear or logistic regression modeling in inferring causal relationships from observational outcome data (Drake and Fisher 1995; Rubin 1997). These advantages have been demonstrated on both theoretical grounds (McLellan, Luborsky, Cacciola, et al. 1985) and in various applications (Rubin 1997; Rubin and Thomas 1992). Although space precludes detailed explication, two of the distinct advantages deserve brief comment (Rubin 1997). First, with propensity scaling it is clearly apparent if the treatment and control groups do not adequately overlap, because controls will be poorly represented in the group characterized by the lowest propensity score. No standard regression model allows this determination.

Second, unlike regression models, the classification approach used in propensity scaling does not rely on any particular functional form of the relationship between potential confounders and the outcome of interest (McLellan, Luborsky, Cacciola, et al. 1985). When multiple potential confounders are being considered, small differences in many covariates can add up to large biases between groups. By taking all of these variables into consideration simultaneously, propensity scale scoring minimizes these risks (Rubin 1997). Propensity scaling thus represents a substantial improvement over multiple regression models in comparing unlike groups and is the best available tool for comparing outcomes in nonexperimental studies (Rosenbaum and Rubin 1983, 1984, 1985; Rubin 1997; Rubin and Thomas 1992).

The major limitation of propensity scaling is that, unlike random assignment in an experimental study, it cannot adjust for factors that are not measured. It is thus an exploratory, not confirmatory analytic procedure in most applications. An additional, more practical condition is that propensity scaling requires large databases so that groups with low propensity to participate in the treatment under study can be adequately represented in the control group. This requirement was readily met by the database used here, and it will be met with increasing frequency as outcomes monitoring becomes standard practice in large healthcare systems.

Hierarchical Modeling

A second important feature of our analysis was the use of hierarchical modeling to take into consideration the similarity of outcomes at particular hospitals. Analysis of these data without consideration of this phenomenon would have led to misleading results because, in this particular data set, a large hospital with the most effective PTSD program also made extensive use of CWT, although patients who participated in CWT at that hospital did no better than patients who did not participate in CWT. Analysis of these data without the use of hierarchical linear modeling would have falsely suggested that CWT is an effective addition to standard PTSD treatment.

The methods presented here thus deserve broad application because they allow rapid, low-cost evaluation of new treatments and methods of service delivery. The availability of such methods may provide a much-needed basis for continued innovation in healthcare systems in which service delivery is closely managed and increasingly standardized.

ACKNOWLEDGMENTS

We would like to thank Thomas Horvath, M.D., Laurent Lehman, M.D., and Paul Errera, M.D. for their administrative leadership of the PTSD outcomes monitoring project; Anthony Campinell, M.S. and Rick Lee for their leadership of VA's CWT program; and Dennis Thompson, Jennifer Cahill, and Don Showalter for data management.

REFERENCES

Blake, D. D. 1994. "Rationale and Development of the Clinician-administered PTSD Scales." *PTSD Research Quarterly* 5 (2): 1-2.

Blumenthal, D. 1996. "Quality of Care: What Is It?" *The New England Journal of Medicine* 335 (12): 891-94.

Brennan, T. A., and D. M. Berwick. 1996. *New Rules: Regulation, Markets and the Quality of American Health Care.* San Francisco: Jossey-Bass.

Bryk, A. S., and S. W. Raudenbush. 1992. *Hierarchical Linear Models.* Newbury Park, CA: Sage Publications.

Clardy, J. A., B. M. Booth, L. G. Smith, C. R. Nordquist, and G. R. Smith. 1998. "Implementing a Statewide Outcomes Management System for Consumers of Public Mental Health Services." *Psychiatric Services* 49 (2): 191-95.

Cohen, J. 1969. *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Davidson, J. R. 1997. "Biological Therapies for Posttraumatic Stress Disorder: An Overview." *Journal of Clinical Psychiatry* 58 (9, Supplement): 29-32.

Drake, C., and L. Fisher. 1995. "Prognostic Models and the Propensity Score." *International Journal of Epidemiology* 24 (1): 183-87.

Ellwood, P. 1988. "Outcomes Management: A Technology of Patient Experience." *The New England Journal of Medicine* 318 (23): 1549-56.

Epstein, A. 1990. "The Outcomes Movement: Will It Get Us Where We Want to Go?" *The New England Journal of Medicine* 323 (4): 265-70.

Fontana, A. F., and R. A. Rosenheck. 1998. "Effects of Compensation-Seeking on Treatment Outcomes Among Veterans with Posttraumatic Stress Disorder." *Journal of Nervous and Mental Disease* 186 (4): 223-30.

———. 1997a. "Effectiveness and Cost of Inpatient Treatment of Posttraumatic Stress Disorder." *American Journal of Psychiatry* 154 (6): 758-65.

———. 1997b. *Outcome Monitoring of VA Specialized Intensive PTSD Programs: FY 1996 Report.* West Haven, CT: Northeast Program Evaluation Center.

———. 1994. "A Short Form of the Mississippi Scale for Measuring Combat-related PTSD." *Journal of Traumatic Stress Studies* 7 (3): 407-14.

Fontana, A. F., R. A. Rosenheck, and H. Spencer. 1993. *The Long Journey Home III: The Third Progress Report on the Specialized PTSD Programs.* West Haven, CT: Northeast Program Evaluation Center.

Friedman, M. J., and R. A. Rosenheck. 1996. "PTSD as a Persistent Mental Illness." In *The Seriously and Persistently Mentally Ill: The State-of-the Art Treatment Handbook,* edited by S. Soreff. Seattle, WA: Hogrefe & Huber.

Johnson, D. R., H. Lubin, M. James, and K. Hale. 1997. "Single Session Effects of Treatment Components Within a Specialized Inpatient PTSD Program." *Journal of Traumatic Stress* 10 (2): 377-90.

Johnson, D. R., R. A. Rosenheck, A. Fontana, H. Lubin, S. Southwick, and D. Charney. 1996. "Outcome of Intensive Inpatient Treatment for Combat-related PTSD." *American Journal of Psychiatry* 6 (5): 771-77.

Kulka, R. A., W. E. Schlenger, J. A. Fairbank, R. L. Hough, B. K. Jordan, C. R. Marmar, and D. S. Weiss. 1990. *Trauma and the Vietnam War Generation: Report* *of Findings from the National Vietnam Veterans Readjustment Study*. New York: Brunner/Mazel.

Lehman, A. F. 1995. "Vocational Rehabilitation in Schizophrenia." *Schizophrenia Bulletin* 21 (4): 645–56.

Lohr, K. N. 1990. *Medicare: A Study of Quality Assurance*. Washington, DC: National Academy Press.

McLellan, A. T., L. Luborsky, J. Cacciola, J. Griffith, F. Evans, H. L. Barr, and C. P. O'Brien. 1985. "New Data from the Addiction Severity Index: Reliability and Validity in Three Centers." *Journal of Nervous and Mental Disease* 173 (3): 412–23.

Millenson, M. L. 1997. *Demanding Medical Excellence: Doctors and Accountability in the Information Age*. Chicago: University of Chicago Press.

Mueser, K., R. E. Drake, and G. R. Bond. 1997. "Recent Advances in Psychiatric Rehabilitation for Patients with Severe Mental Illness." *Harvard Review of Psychiatry* 5 (3): 123–39.

National Committee on Quality Assurance. 1995. *National Committee on Quality Assurance Report Card Pilot Project*. Washington, DC: NCQA.

Relman, A. 1988. "Assessment and Accountability: The Third Revolution in Health Care." *The New England Journal of Medicine* 319 (18): 1221–22.

Rosenbaum, P. R., and D. B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–38.

———. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (6): 516–24.

———. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rosenheck, R. A. 1996. *Department of Veterans Affairs National Mental Health Program Performance Monitoring System: Fiscal 1995 Report*. West Haven, CT: Northeast Program Evaluation Center.

Rosenheck, R. A., and A. F. Fontana. 1996. "Treatment of Veterans Severely Impaired by PTSD." In *Emotional Aftermath of the Persian Gulf War*, edited by R. J. Ursano and A. E. Norwood. Washington, DC: American Psychiatric Press.

Rosenheck, R. A., A. F. Fontana, H. Spencer, and S. Gray. 1997. *Long Journey Home V: Treatment of Posttraumatic Stress Disorder in the Department of Veterans Affairs. Fiscal Year 1996 Service Delivery and Performance*. West Haven, CT: Northeast Program Evaluation Center.

Rosenheck, R. A., N. Wilson, and M. Meterko. 1997. "Consumer Satisfaction with Inpatient Mental Health Treatment: Influence of Patient and Hospital Factors." *Psychiatric Services* 48 (12): 1553–61.

Rubin, D. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8, Supplement): 757–63.

Rubin, D., and N. Thomas. 1992. "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Covariates." *Biometrika* 79 (6): 797–809.

Sederer, L. I., B. Dickey, and R. C. Hermann. 1996. "The Imperative of Outcomes Assessment in Psychiatry." In *Outcomes Assessment in Clinical Practice*, edited by L. I. Sederer and B. Dickey, pp. 1–7. Baltimore, MD: Williams and Wilkins.

Seibyl, C. L., R. A. Rosenheck, L. Corwel, and S. Medak. 1997. *Compensated Work Therapy: Veterans Industries Fiscal Year 1996.* West Haven, CT: Northeast Program Evaluation Center.

Solomon, S., E. T. Gerrity, and A. M. Muff. 1992. "Efficacy of Treatments for Post-traumatic Stress Disorder." *Journal of the American Medical Association* 268 (5): 633–38.

Spitzer, R. L., and J. B. W. Williams. 1985. *Structured Clinical Interview for DSM-III PTSD.* New York: New York State Psychiatric Institute.

Steinwachs, D. M., A. W. Wu, and E. A. Skinner. 1994. "How Will Outcomes Management Work?" *Health Affairs* 13 (1): 153–62.

Weathers, F. W., and B. T. Litz. 1994. "Psychometric Properties of the Clinician-administered PTSD Scale, CAPS-1." *PTSD Research Quarterly* 5 (2): 2–6.